Task complexity, language proficiency and working memory: Interaction effects on second language speech performance

Article

Accepted Version

Task complexity, language proficiency and working memory: Interaction effects on second language speech performance

Abstract: With the aim of developing a more reliable understanding of the effects of task complexity and learner-internal factors on L2 performance, a 2 × 2 within-between participant study was designed to examine the effects intentional reasoning has on L2 performance, and whether learner language proficiency and working memory mediates these effects. Forty- eight learners of English performed two video-based narrative tasks of varying degrees of intentional reasoning, after taking Oxford Placement Test, Elicited Imitation Tasks and backward-digit span tasks. The results demonstrate that intentional reasoning had significant effects on complexity and accuracy, but no impact on fluency. Regression analyses indicated that proficiency and working memory reliably predicted accuracy across both task types. However, language proficiency and working memory contributed differentially to models predicting lexical complexity and speed fluency in the two task types, highlighting the interaction between task complexity and learner-internal factors. Keywords: second language speech performance, task complexity, intentional reasoning, language proficiency, working memory.

Introduction

Research in task-based language teaching (TBLT) over the past decades has witnessed a growing interest in conceptualising, defining and investigating cognitive task complexity (TC) (e.g., Awwad, Tavakoli & Wright, 2017; Tavakoli & Foster, 2008; Cho, 2018; Declerck & Kormos, 2012; Robinson, 2007; Sasayama, 2016). TC, defined as "attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (Robinson, 2001, p. 29), or more simply as "the cognitive load of a second language (L2) communication task" (Sasayama, 2016: 231), is central to research in both TBLT

and second language acquisition (SLA) as it is assumed to affect L2 processing, production and acquisition. The interest in researching TC is inspired by theoretical and methodological questions such as whether TC can facilitate L2 production and acquisition, and how it interacts with psycholinguistic processes of attention allocation, noticing, and automaticity. From a pedagogic perspective, the interest in TC is rooted in the need for developing an index of complexity to be used in task design and task sequencing in language teaching, syllabus design and assessment (Malicka, 2014; Robinson, 2015). Despite the substantial research interest in TC, identification of design features and variables that contribute to TC remains a challenge. Several variables have been proposed and examined, e.g., task structure and storyline complexity (Tavakoli & Foster 2008; Tavakoli & Skehan 2005), still it appears that there are many more yet to be investigated.

In a systematic review of the literature on TC, Jackson and Suethanapornkul (2013) identified two key limitations to TBLT research: a paucity of research into various aspects of TC, e.g. reasoning demands, as well as a lack of consistency in the operationalization of these variables. Yet another important, and relatively neglected focus in TBLT research is the relationship between TC and learner-internal variables. Recent TC research (Gilabert & Munoz, 2010; Kormos & Trebits, 2011) has presented evidence that the impact of learner-internal variables, e.g. language proficiency (LP) and working memory (WM), on task performance is weak relative to the effects of TC. However, the relationship between TC and the learner-internal variables remains an insufficiently researched focus, despite the fact that investigating this relationship is believed to be a promising path towards developing a more in-depth understanding of the way TC may mediate L2 production and acquisition (Declerck & Kormos, 2012; Gilabert & Muñoz, 2010; Malicka & Levkina, 2012; Révész, 2011). Our study is an attempt to fill some of these gaps by investigating an under-researched TC variable, i.e., intentional reasoning (IR), and its relationship to some individual learner differences. We aim

to provide not only a more systematic approach to defining and operationalizing TC, but also a further insight into the ways two important learner variables, i.e., LP and WM, may mediate the effects of TC. While prior TC literature has mainly examined LP and WM in isolation, the gap that the current study aims to help fill is considering the effects of both variables and the possible interaction between the two on task performance.

Literature Review

Task complexity

There is little disagreement among TBLT researchers that TC is a complex and multidimensional construct (Robinson, 2007; Sasayama, 2016; Skehan, 1998, 2014; Vasylets, Gilabert & Manchon, 2017) interacting not only with task material, design and mode, but with learner cognition and individual differences (Révész, 2011; Robinson, 2011). Liu and Li (2012) argued that, when defining TC broadly, researchers refer to three distinct qualities of a task: a) its structure (e.g., number of elements and characters), b) its resource requirements (e.g., what is needed to perform the task), and c) interaction between a task and learner variables (e.g., cognitions and WM). Many TBLT researchers, however, have sought a more detailed and theoretically supported framework for defining TC, and situate their studies with reference to two influential cognitive-interactionist TC models, i.e. Limited Attentional Capacity (Skehan, 1998, 2015) and the Cognition Hypothesis (Robinson, 2007, 2015).

Drawing on a multiple-resource model of attention, the Cognition Hypothesis proposes that the human brain has access to a pool of multiple resources, and that therefore increasing TC encourages access to multiple resources promoting more complex and accurate language production. It also theorizes that this higher cognitive demand both creates appropriate opportunities for learning and facilitates L2 acquisition. By contrast, Skehan's Limited Attentional Capacity model (2015) highlights the limited nature of the learner's processing

capacity, and assumes that a higher cognitive demand requires greater attentional resources, thus forcing learners to prioritise their allocation of attention. This often results in a competition between different performance dimensions, especially between form (i.e., complexity and accuracy) and meaning (i.e., fluency), or between different aspects of form. The latter model envisages that the pressure resulting from high cognitive load will limit opportunities for development and acquisition. While we draw on this body of literature in our research overview, we do not aim to match our findings against either model. Instead, we aim to discuss TC in a broader perspective of individual learner variables. More central to our study is the fact that TBLT research has so far repeatedly examined variables such as Here-and-Now versus There-and-Then, task structure and planning time, without sufficiently examining other aspects of TC. IR is one such TC variable.

Intentional reasoning

Robinson (2007; 2015) proposes that tasks that require describing motion events (spatial reasoning), explaining reasons for actions (causal reasoning), and reading other peoples' minds (intentional reasoning) have high IR demands, and therefore lead learners' attention to using more accurate and complex language to convey such demands. Robinson (2007) argues that a task that requires IR encourages L2 learners to adopt complex linguistic structures (e.g., subordinating conjunctions), to create cohesion between intentions, actions and predictions, and to use a lexis with higher complexity (e.g., mental states verbs, adverbs of uncertainty) that allows these intentions to be described. On a less positive side, however, IR is expected to reduce fluency.

To the best of our knowledge, there are only two studies that have examined the impact of IR on speech performance. Robinson (2007) used a continuum of IR through three picture sequencing and picture telling tasks performed by 42 Japanese students. IR was operationalised by the demand to explain the intentions of each story character. The simplest task entailed

explaining one character's IR, whereas the more complex tasks involved more characters whose intentions were reliant on others' ideas and desires. The findings were not in line with the predictions of the Cognition Hypothesis as increasing IR demands did not elicit more complex and accurate language or lead to reduced fluency. Ishikawa (2008) was the second study to investigate the effect of manipulating TC through IR demands. In this study, IR was operationalised as the demand to explain changes in human relationships at workplace. The 24 participants performed three tasks that required them to report changes in relationships between staff, based on several trouble triggers. Three levels of no reasoning, simple reasoning, and complex reasoning were designed. Similar to Robinson's (2007) study, the extra IR demand was operationalised by Ishikawa (2008) as the number of characters involved in the tasks, i.e. the complex versions had more characters and hence needed more reasoning. The no reasoning task required only describing current relationships between the characters, the simple reasoning entailed explaining changes in relationships of two characters, and the more complex reasoning task involved four characters whose intentions were reliant on the others' ideas and desires. Supporting the CH predictions, the findings showed that higher IR demands produced higher complexity and higher accuracy, whereas fluency performance decreased. Notwithstanding the value of the findings of these studies, a methodological limitation in their design makes it difficult to interpret the results. In both studies, IR seems to conflate the number of characters/elements in a task and the amount of reasoning required, i.e., the more complex task has more characters, concepts and elements and as such needs more IR for each character. The interdependence of these aspects of task design is an important issue that researchers such as de Jong and Vercelloti (2015) and Sasayama (2016) have warned fellow researchers against. There are at least two ways of manipulating TC regarding IR demands to avoid the interdependence issue. First, TC can be operationalized in terms of the number of elements in the task with the more complex task having more elements. Second, TC can be operationalized

in terms of the amount and level of reasoning required to complete the task. The current study focuses on the latter option.

Drawing on the Cognitive Psychology literature, Awwad et al. (2017) consider IR as "a critical element involved in a) observing others' actions and behaviours, and b) arriving at conclusions about others' thoughts, intentions and beliefs" (Awwad et al. 2017: 161). IR, which involves hypothesizing, interpreting and drawing conclusions about others' thoughts, actions and behaviours, is a serial cognitive process (Leighton, 2004) dependent on a chain of logical premises and hypotheses (Gilhooly, 2004). The serial processes and semantic operations that are required to create this logical chain add to the cognitive load by placing an extra burden on attention and working memory, especially while generating idea units and constructing an appropriate preverbal message (Levelt, 1989). From a linguistic point of view, it can be argued that describing these thoughts and intentions, as well as explaining and justifying them, would invite use of specific language that denotes intentionality and reasoning. We have argued (Awwad et al. 2017) that representing IR in the English language is expected to encourage syntactically more complex structures (e.g., logical subordinators), and lexically more complex words (e.g., cognitive status verbs). The use of hypothetical language of a formulaic nature (e.g., I think and I suppose) would also promote accuracy, at least at the level of short clauses. Testing the aforementioned predictions, our results (Awwad et al. 2017) of examining speech performance of L2 learners on tasks with different degrees of IR, showed that speech performance in the +IR condition was associated with syntactically more complex and accurate but, surprisingly, less lexically diverse language. No statistically significant differences were observed in fluency of the learners' performances across the two conditions. One way to explain the unanticipated results regarding lexical diversity and fluency, was to hypothesize that TC had interacted with the learners' individual differences in LP and WM (see Awwad et al. 2017 for further details). This post-hoc observation was the point of departure for the current study.

Language proficiency and task complexity

Investigating a complex and abstract process such as TC becomes more intricate when learners are performing tasks in a language in which they are not proficient. LP, i.e., "the linguistic knowledge and skills that underlie L2 learners' successful comprehension and production of the target language" (Gaillard & Tremblay, 2016, p.420), is assumed to play a major role in L2 processing that depends on conscious and controlled attention (Kormos, 2011), and it can either drive or hinder language performance by the amount of automaticity available for language encoding and attention allocation during performance. Research on the interaction between TC and LP attempts to explain whether variations in LP lead to variations in allocating attention, controlling learner interlanguage, and monitoring speech production. By maintaining a smooth flow of linguistic resources during performance, higher levels of LP are argued to support L2 processing by assisting learners to engage in parallel processing, and freeing up attentional resources to attend to different aspects of performance (Kormos, 2011). Therefore, an interaction between LP and TC is expected in that high-proficiency learners are likely to operate more successfully while processing and performing cognitively demanding L2 tasks.

Given TC as a research focus, only a small number of studies have investigated the interaction between LP and TC (e.g., Ishikawa, 2006; Kuiken & Vedder, 2008; Malicka & Levkina, 2012). Employing X-lex and Y-lex vocabulary size tests and the Oxford Placement Test to measure the participants' LP, Malicka and Levkina (2012) investigated whether LP regulated the effects of reasoning demands and the number of elements in instruction-giving tasks. Their results suggested that the high-proficiency group produced more complex and accurate performance on the complex tasks, whereas the low-proficiency group produced more fluent language. Investigating the interaction between LP and +/-Here and Now on L2 learners' narrative writing, Ishikawa (2006) found that TC influenced all aspects of performance except for lexis. Using a Cloze test to assess LP, Kuiken and Vedder (2008) examined the interaction between

the number of elements in a task and LP on L2 learners' writing performance. Although their findings revealed major effects of LP on grammatical complexity, accuracy and lexis, no interaction effects were found between +/- Here and Now and LP on any aspect of writing performance. Two conclusions can be reached based on the mixed findings reported above. First, it is possible to link the mixed results with the different aspects of TC, task modes and types used in these studies. Another way to interpret these results is to highlight the lack of consistency in assessing LP in these studies. In the current study, we are keen to investigate LP from a broader perspective by assessing proficiency in terms of both explicit and implicit knowledge. SLA research (DeKeyser, 2003, 2009; Ellis, 2009; Hulstijn, 2005) has postulated that language proficiency is comprised of two different underlying constructs, i.e. implicit and explicit knowledge, characterised by presence or absence of conscious awareness. Drawing on unconscious and intuitive knowledge, implicit knowledge is procedural in nature and results in fluent speech production. Explicit knowledge, on the other hand, draws on declarative knowledge and is mainly acquired through conscious awareness, and therefore assumed to result in controlled processing (DeKeyser, 2003; Hulstijn, 2005). Despite the significance of these two types of knowledge, most TC studies have examined proficiency in the explicit type only (e.g., Declerck, & Kormos, 2012; Malicka & Levkina, 2012). Given that previous research in this area has failed to examine the differential contributions of explicit and implicit knowledge to task performance, the current study aims to explore the interaction between TC and LP in relation to both implicit and explicit knowledge.

Working memory and task complexity

The second learner-internal factor we are examining here is WM. WM, "a multi-component system which is responsible for active maintenance of information in face of ongoing processes and/or distraction" (Conway et al., (2005, p. 770), is envisaged to interact with L2 performance and development in general and with TC in particular (Cho, 2018; Gilabert & Muñoz, 2010;

Kormos & Trebits, 2011; Mitchell, Jarvis, O'Malley & Konstantinova, 2015; Mota, 2003). The tendency to incorporate WM in TC studies stems from the notion that WM is at stake in the performance of L2 complex tasks (Kyllonen & Christal, 1990) for its potential influence on regulating L2 learners' linguistic repertoire and attentional resources during language performance (Wen, Mota, & McNeill, 2015). While several studies have investigated the relationship between WM and L2 acquisition, not many have examined the relationship between TC, LP and WM. In a correlational study, Gilabert and Munoz (2010) explored whether variation in WM and LP would explain variation in L2 performance. A reading span test measured the participants' WM, while three tests assessed LP, i.e. the Oxford Placement Test, vocabulary size tasks, and a phonetic classification task. The 59 participants were allocated to low-high LP groups, and performed only one video-based narrative task. Though Gilabert and Munoz did not find any correlation between WM and LP, they did find that WM correlated with lexical complexity and fluency. Moreover, LP was found to correlate with all aspects of speech performance except syntactic complexity. LP was found to be a reliable predictor of lexis, whereas WM was not.

Kormos and Trebits (2011) investigated the relationship between WM and TC on L2 oral performance. A backward-digit span task measured the participants' WM. The participants performed two narrative tasks with varying TC, i.e. telling a story (simple) and inventing a story (complex). It was found that high WM benefited syntactic complexity only in the simple task in terms of ratio of subordination and length of clause. The complex task elicited performances of more accuracy and less lexical complexity, but no effect was observed for grammatical complexity or fluency. It is necessary to mention that this study did not employ a standardised LP test, but instead relied on teachers' judgments of student LP. Using a standardised LP test, Mitchell et al. (2015) investigated the LP-WM interaction effects on L2 processing and development. Based on their TOFEL scores, 36 Chinese learners of English were grouped into

three levels of proficiency (beginner, intermediate, advanced). Their WM was measured by an operation span task and forward-digit span tasks in L1 and L2. To measure the participants' proficiency, the authors used elicited imitation and reading tasks in English. While the findings did not show any correlations between LP scores and L1 digit span and operation span scores, LP did correlate with the L2 digit span scores. The study found a stronger relationship between WM and LP in the case of high proficiency learners.

The current study

This study aims to investigate the effects of TC on L2 speech performance across different LP and WM levels. It is an attempt to examine to what extent LP (both implicit and explicit L2 knowledge) and WM can predict performance on tasks requiring different levels of IR. The research questions guiding the current study are:

RQ1: What are the effects of TC, operationalized in terms of the amount of IR required in a task, on learners' L2 speech performance, measured by syntactic complexity, lexical complexity, accuracy, and fluency?

RQ.2: To what extent can LP and WM predict learner performance on tasks of different degrees of IR?

Methodology

In a within-between-participants factorial design, 48 L2 learners of English performed two video-based oral narrative tasks with different levels of IR. The order of performing the tasks was counterbalanced to control for any possible practice or order effect. IR was a within-participant variable with two levels, i.e., the task with more IR demand (+IR) and the one with less IR demand (–IR) in this paper. It is necessary to note that we consider IR spanning over a continuum, and therefore the use of + and – does not denote a dichotomy. LP and WM were

continuous between-participant variables. The dependent variables were syntactic and lexical complexity, accuracy and fluency of learners' L2 performance.

Participants

The participants were 48 students at a private secondary school in Jordan. Since the school was a single-sex school, all the participants were males. They were aged 16, with Arabic as their first language. The demographic data showed they had very similar schooling and language instruction experiences, i.e., they had studied English for about ten years at school, and had never lived in an English-speaking country. The school where data were collected teaches the national Jordanian syllabus providing one English lesson every day using textbooks that corresponded to different CEFR levels (CEFR, 2001). Using an internal placement test, the school had grouped the students into three levels (A, B, C) according to their English proficiency level. The students were then assessed throughout the academic year on the school internal tests and portfolios of continuous assessments. Before the data collection, we sent out an invitation to all students in year X in this school asking for volunteers to participate in the study. The data were collected from 52 participants, but due to a technical problem with the audio recording, we had to remove data from four of the participants.

Language proficiency test

Despite the abundant research interest in investigating spoken proficiency as a key construct of L2 ability, there is some evidence to suggests that operationalisation, measurement and analysis of spoken proficiency is not always done carefully and systematically, inevitably resulting in poor test reliability and/or validity (Bachman, 1990; Fulcher, 2014). This limitation has, for instance, been reflected in studies that use a pen-and-paper test to represent speaking proficiency, or when a multiple-choice grammar test is used to investigate learner communicative adequacy. In line with this debate, Leal (2018) argues that measurement and analysis of proficiency in such a limited manner will have inadvertent consequences for

research findings both theoretically and empirically. To prevent such negative effects, Leal (2018) proposes that researchers should consider and analyse proficiency in its full sense and as a continuous variable to show variance among the participants.

As discussed earlier, the limited approach to assessing proficiency in previous studies that examined TC across different levels of LP is a source of ambiguity in understanding and interpreting the results of the studies summarized above. To make up for such limitations and to develop a more in-depth insight into the effects of LP on TC, two tests were used to investigate explicit and implicit knowledge. The Oxford Placement Test (OPT) (Alan, 2004) is assumed to measure learners' L2 explicit knowledge, and elicited imitation tasks (EIT) (Wu & Ortega, 2013) are supposed to assess their implicit knowledge (Ellis, 2009; Erlam, 2006). We used a pen-and-paper version of the OPT containing 60 multiple-choice questions to assess their explicit knowledge (a maximum score of 60). The version of EIT used in the study comprised ten sentences increasing in number of syllables from 8 to 19 (Wu & Ortega, 2013; Yan, et.al., 2015). This version was chosen because previous research has provided evidence that as sentences gradually increase in length, test takers feel under pressure to access their interlanguage to produce the sentences. As the increase in task demands encourages use of implicit knowledge, the process allows researchers to learn more about how L2 learners' implicit linguistic knowledge is activated (Erlam, 2006; Wu & Ortega, 2013; Yan, et.al., 2015). Based on the accuracy of the imitation, each sentence was given 0-4 points with 40 points as the maximum score. As for OPT, since it was a multiple-choice test, a second marker was employed to cross-check the accuracy of the marking and no disagreement was found. As for EIT, a second rater checked the accuracy of the rating. Pearson correlation coefficient of 92% was achieved between the researchers and the rater.

Working memory test

SLA research suggests that backward-digit span tests are appropriate tools to measure L2 learners' WM as they are language independent and therefore minimise any impact of proficiency on WM scores (Harrington & Sawyer, 1992), and allow for both storage and processing to be examined (Kormos & Trebits, 2011; Mitchel et al. 2015). In this study, we used backward-digit span tests in both L1 (Arabic) and L2 (English) to cross-check that they were language independent. Designed by the researchers, the WM tests comprised seven sets of increasing numbers, i.e. 3-9. The sets were audio recorded by one of the researchers at one digit per second. The participants were required to listen to these sets and repeat them backwards. The Arabic and English versions were counterbalanced between participants. The participants were given three attempts for each set. Each participant's WM span was determined based on the last set of digits he repeated successfully twice. That is if a participant failed to repeat two sets out of three of the same span, his WM span would be the last set he repeated successfully twice. The participants' WM span scores in the two tests ranged between 4 and 9. Considering the strong correlation between L1 and L2 WM tests in this study ($r = .87$, $p = .001$), the L1 test scores are used in the analysis. The descriptive statistics for OPT, EIT and WM scores are summarised in Table 1 below.

Table 1. Descriptive statistics for LP and WM scores

Test	Min.	Max.	Mean	SD
Oxford Placement Test	20	50	34.1	7.00
Elicited Imitation Task	19	40	28.3	4.86
Backward-digit WM Test	4	9	5.16	1.22

N = 48

The video tasks

For comparability purposes, the same tasks as those in our previous study were used (Awwad et al. 2017). The video-based tasks were adopted from Pat & Mat (Beneš & Jiránek, 1976), an animated cartoon series about two friends who deal with everyday challenges and troubles in optimistic, creative and funny ways. de Jong and Vercelloti's (2015) framework was used to choose the video clips. Based on this framework, a number of factors were carefully considered to ensure that the two clips were similar with respect to the number of characters and elements, duration and storyline. However, the two clips differed in terms of the actions involved and the amount of IR needed to justify the characters' actions. Besides controlling for IR at content level, we operationalized it at task instruction level. While the instructions in the -IR task asked the participants to tell and describe the story, the +IR task encouraged the participants not only to tell and describe the events, but also to read the characters' thoughts and intentions and to predict and explain their decisions, actions and reactions. The duration of each video clip was 120 seconds. The choice of tasks was validated through a retrospective questionnaire in which the participants rated the -IR and +IR tasks in terms of the degree of difficulty on a four-point scale. The participants significantly rated the +IR task as more difficult compared to the -IR task (t = -7.43, p = .00, d = -1.52) suggesting that the task that required more IR was assumed to be more complex as it was designed to be (for further details see Awwad et al. 2017).

The participants met with one of the researchers in a quiet room. To avoid test fatigue, they took the LP and WM tests in one room and performed the tasks in another. Pre-task planning time was not provided to ensure that the relationship between TC, WM and LP was not mediated by strategic planning. However, a demo video clip (30 seconds) was shown to the participants to familiarize them with the task type. The researcher read the instructions in the participants' L1 and L2 and then asked them to narrate the story to him in their L2, i.e. English. A digital voice recorder with a headphone was used to record the participants' performances.

At the end of the video clips, each participant was given an extra 20 seconds to finish their performance if needed.

Data coding and analysis

Using SoundScriber software (Breck, 1998), the data were transcribed and coded for measures of syntactic and lexical complexity, accuracy, and fluency. AS-unit (Foster, Tonkyn & Wigglesworth, 2000) was employed to segment the transcriptions into units of analysis. Following the literature in this area (e.g. Kuiken & Vedder, 2011; Wang & Skehan, 2014), three measures of syntactic complexity were included: mean length of AS-unit, mean length of clause, and ratio of subordination. The choice of these measures was justified by the need to incorporate both length and subordination measures for a reliable exploration of syntactic complexity at higher levels of proficiency (Norris and Ortega, 2009). The choice of syntactic complexity measures also takes into account the recommendations of Inoue (2016) who suggests that "researchers need to consider seriously the task-essentialness of subordinate clauses when deciding on the tasks to use for research" (p.495). Lexical complexity was measured in lexical sophistication (PLex Lambda) and lexical diversity (D). PLex Lambda is a measure of sophistication that assesses the occurrence of less frequent words in a text evaluating knowledge of more sophisticated words and therefore is a more reliable measure of sophistication in short texts (Meara & Miralpeix, 2016). PLex Lambda was calculated using *Lognostics Toolbox*, a free software that offers different tools for researching vocabulary, including sophistication (Meara & Bell, 2001). D is a corrected type-token ratio measure that responds to variations in text length (Malvern & Richards, 2002). D was calculated by the Voc-D function available in *Coh-Metrix software* (Graesser et al., 2003).

Drawing on the existing evidence about the robustness of global measures of accuracy (Ellis & Barkhuzein, 2005; Ong & Zhang, 2010), we chose two global measures to represent accuracy.

The first measure is percentage of error-free clauses (EFC), which is shown to be sensitive to detecting accuracy across different levels of LP (Ellis & Barkhuzein, 2005). An error-free clause, in our analysis, is one in which there are no errors in terms of grammar, word choice or language use. Error-free clause is calculated by dividing the number of error-free clauses by the total number of clauses produced in a performance, multiplied by 100.

The second measure used in this study is a more recently developed measure of accuracy in which errors are examined in terms of their seriousness. Foster and Wigglesworth (2016) have argued that error-free clause may fail to distinguish between errors of different gravity, and as such they have proposed a more systematic approach to measuring clause-level accuracy. Foster and Wigglesworth (2016: 98) argue that a weighted clause ratio (WCR) is a more appropriate measure of global accuracy as it "classifies errors at different levels" and distinguishes between "those that seriously impede communication, those that impair communication to some degree, and those that do not impair communication at all". We are keen to find out whether WCR is more sensitive than percentage of EFC in detecting differences across different LPs. Using both measures would also enable us to examine the possible relationship between the two.

Following fluency research literature (Skehan, 2003; Tavakoli, Campbell & McCormack, 2016; Tavakoli, Nakatsuhara & Hunter 2017), we chose four measures to represent speed, breakdown and repair fluency. For speed fluency, we chose pruned speech rate as it is suggested to be a reliable measure of global speed fluency in L2 research (Kahng, 2014; Segalowitz, 2010; Tavakoli & Skehan 2005). Given the robust evidence in SLA research about the usefulness of mean length of silent pauses at mid-clause and end-clause positions, these two measures were selected to represent breakdown fluency (Kahng, 2014). To characterize repair fluency, a global measure of repair that included all repair types of repetition, hesitation, reformulation, replacement, and false start was used. This measure is commonly believed (e.g., Lennon, 1990; Skehan, 2003, Foster & Tavakoli 2009) to represent the amount of repair during speech

production. A threshold of > 0.40 second was used to distinguish silent pauses (Tavakoli & Skehan 2005). Temporal measures were calculated using PRAAT (Boersma & Weenink, 2008). All measures of fluency were calculated per 60 seconds. To check the reliability of data coding, a second rater checked 20% of the transcriptions. Measures of accuracy were coded using the researchers' judgement, and they were further cross-checked by an English native speaker language expert. Pearson correlation coefficient revealed high agreement between the researchers and the raters with respect to the measures of complexity (94%), accuracy (89%) and fluency (91%). The high inter-rater reliability achieved allowed us to proceed with data analysis.

Results

A MANOVA was run to identify whether there were statistically significant differences between performances on the two tasks (-IR and +IR) across different dependent variables. Following previous research in this area (e.g., Kahng, 2014; Tavakoli & Skehan, 2005; Skehan & Foster, 2012), the most consistent CALF measures were used to represent the four aspects of performance in the MANOVA. The measures were mean length of AS unit, D, percentage of error-free clauses, and pruned speech rate. All required assumptions were checked prior to the analysis and no violations were detected including normality, homogeneity and linearity. Cohen *d* effect size was calculated when significant differences were obtained (Cohen, 2013). However, Plonsky and Oswald's (2014) interpretation of effect size was adopted, i.e., small (0.4), medium (.70) and large (1.00).

The MANOVA output revealed a statistically significant difference with a large effect size for the four dependent variables combined (Wilks' Lambda = .291; $F = 26.77$, $p = .000$; $\eta^2 = .709$). When considering each dependent variable separately, the differences were also statistically significant in terms of *syntactic complexity* (Wilks' Lambda = .510; $F = 45.22$, $p = .000$; η^2

= .490), *lexical complexity* (Wilks' Lambda = .913; F = 4.46, p = .04; η^2 = .087), *accuracy* (Wilks' Lambda = .386; F = 74.67, p = .000; η^2 = .614), and *fluency* (Wilks' Lambda = .605; F = 30.71, p = .000; η^2 = .395). The results of the MANOVA, providing evidence of significant effects on L2 performance, were followed by paired-sample t-tests to answer investigate the effects of IR demands on L2 learners' speech performance. The t-tests results are summarised in Table 2 below.

Table 2. T-tests results for +IR and -IR task performances

Aspects	**Measures**	- IR	+ IR	t-test	Sig. (2-tailed)	Effect size
		Mean (SD)	**Mean (SD)**	***t***	***p***	***d***
Syntactic Complexity	Mean length of AS unit	**6.77 (1.23)**	**7.56 (1.06)**	**-5.21**	**.000***	**.69**
	Mean length of clauses	5.21 (.54)	5.12 (.43)	1.37	.177	.18
	Ratio of subordination	**1.29 (.16)**	**1.47 (.17)**	**-6.72**	**.000***	**1.09**
Lexical Complexity	Lexical diversity (D)	**25.04 (10.27)**	**23.14 (8.83)**	**2.11**	**.040***	**.20**
	Lexical sophistication (PLex)	**1.18 (.34)**	**.85 (.26)**	**6.17**	**.000***	**1.09**
Accuracy	Error free clauses	**43.25 (17.04)**	**57.72 (15.99)**	**-8.64**	**.000***	**.88**
	Weighted clause ratio	**.79 (.08)**	**.85 (.06)**	**-5.76**	**.000***	**.85**
Fluency	Pruned speech rate	**92.1 (23.49)**	**103.4 (26.87)**	**-5.54**	**.000***	**.45**
	Mean length of mid-clause silent pauses	.94 (.34)	.85 (.28)	1.61	.114	.29
	Mean length of end-clause silent pauses	1.30 (.49)	1.20 (.50)	1.77	.082	.20
	Number of repairs	8.92 (4.05)	9.85 (4.93)	-1.62	.111	.20

df = 47, *p (2-tailed) < 0.05

Effects of IR on L2 speech performance

The results of the t-test for syntactic complexity showed that in the +IR task participants produced statistically longer AS-units (t = 5.21, p = .000, d = .69) a higher *ratio of subordination* (t = -6.72, p = .000, d = 1.09) with a large effect size. Regarding mean length of clauses, although performances in the -IR task generated longer clauses (M = 5.21, SD = .54) than the +IR task (M = 5.12, SD = .43), the difference did not reach a significant level. For lexical complexity, the findings revealed that performance in the +IR task was *lexically* less *diverse* when compared to the -IR task with a statistically significant result and a small effect size (t = 2.11, p = .40, d = .20). Finally, language performance in the +IR task was characterized by less *lexical sophistication* measured by PLex Lambda, reaching a significant difference level and a large effect size (t = 6.17, p = .000, d = 1.09).

Regarding accuracy, performance in the +IR task elicited a higher *percentage of error free clauses* with a medium effect size (t = -8.64, p = .000, d = .88) and a higher ratio of weighted clauses (t = -5.76, p = .000, d = .85). As for fluency, the +IR task elicited a significantly higher *speech rate* with a small effect size (t = 5.54, p = .000, d = .45). Regarding mean length of silent pauses, although performances in the -IR task generated longer pauses mid-clause (M = .94, SD = .34) and end-clause (M = 1.30, SD = .49) than in the +IR task (M = .85, SD = .28), (M = 1.20, SD = .50), the differences did not reach significant levels. *Number of repairs* showed that the participants produced more repairs while performing the +IR task (M = 9.85, SD = 4.93) compared to the -IR task (M = 8.92, SD = 4.05), but this difference was not significant.

Language proficiency and working memory predicting language performance

Research Question 2 asked whether LP and WM were reliable predictors of L2 speech performance on tasks of varying levels of TC. Multiple regression analyses were performed with LP and WM as predictor factors, where composite measures of syntactic complexity,

lexical complexity, accuracy and speed fluency were employed as dependent variables. Separate analyses were run for +IR and –IR task performances.

The non-significant correlation between WM and explicit L2 knowledge (r = .028, p = .426) and implicit L2 knowledge (r = .016, p = .457) indicated that the two predictors of LP and WM tapped into different aspects of language performance, and therefore were assumed to be suitable for inclusion in the regression analysis. Rather predictably, a significant correlation was found between explicit and implicit L2 knowledge (r = .65, p = .001).

Table 3. Multiple regressions for LP and WM as predictors of performance in **-IR** task

Outcomes	**Predictors**	**Correlations**		**Regression models**			**Coefficients**				
DVs	**IVs**	***r***	***p***	***F***	***p***	***R²***	***B***	***SE***	***β***	***t***	***p***
Syntactic complexity	EXPLP	-.165	.132	.664	.591	.042	-.056	.041	-.269	-1.38	.174
	IMPLP	-.014	.461				.048	.059	.161	.826	.413
	WM	-.015	.461				.006	.177	.005	.033	.974
Lexical complexity	EXPLP	-.203	.083	.820	.490	.053	-.298	.245	-.235	-1.21	.232
	IMPLP	-.099	.252				.096	.354	.053	.272	.787
	WM	.106	.237				.714	1.06	.098	.670	.506
Accuracy	EXPLP	**-.370**	**.005***	**3.72**	**.018***	**.202**	-.627	.406	-.274	-1.54	.130
	IMPLP	**-.312**	**.016***				-.452	.587	-.137	-.771	.445
	WM	**.244**	**.047***				3.12	1.76	.239	1.77	.083
Speed fluency	EXPLP	**-.329**	**.011***	2.29	.091	.135	-2.00	1.40	-.264	-1.42	.160
	IMPLP	**-.262**	**.036***				-1.01	2.02	.093	-.502	.618
	WM	.158	.141				6.63	6.10	.153	1.088	.283
Breakdown fluency	EXPLP	-.063	.335	1.08	.367	.069	-.351	.256	-.263	-1.37	.177
	IMPLP	.129	.191				.582	.370	.302	1.57	.122
	WM	-.109	.231				-.923	1.11	-.121	-.829	.411
Repair fluency	EXPLP	.109	.231	.672	.574	.044	.265	.214	.240	1.23	.223
	IMPLP	-,51	.366				-.328	.309	-.206	-1.05	.634
	WM	-.091	.270				-.509	.931	-.081	-.547	.587

*p < 0.05, df (3, 44)

As shown in Table 3, the results regarding whether LP and WM predict performance in the -IR task revealed that the regression model was not significant for *syntactic complexity* (F (3, 44) = .884, p = .591), suggesting that the syntactic complexity of learners' performance could not be explained by variations in their explicit and implicit LP or WM. As for *lexical complexity,* the non-significant regression (F (3, 44) = .820, p = .490) also indicated that neither explicit and implicit LP nor WM could be considered as reliable predictors of learners' lexical complexity.

The regression model for *accuracy* reached a significant level (F (3, 44) = 3.72, p = .018), explaining 20% of the variance. All three predictors, i.e. explicit LP (p = .005), implicit LP (p = .016), and WM (p = .047) contributed significantly to the model. For *speed fluency,* although the regression model failed to reach a statistically significance level (F (3, 44) = 2.29, p = .091), a 14% of the variance in this performance was explained by explicit and implicit LP knowledge (p = .011 and p = .036 respectively). WM did not make a contribution to this model (p = .141). In sum, the results of the multiple regression analyses regarding -IR performance showed that only the model for accuracy was statistically significant, whereas the model for speed fluency showed signs of approaching a statistically significant level. As for the individual contributions, only explicit and implicit LP contributed significantly to both models, whereas WM contribution reached a significant level for only the accuracy model. These results suggested that LP and WM were reliable predictors of accuracy in terms of L2 performance in the -IR task, whereas LP could to some extent predict speed fluency only as well.

Turning to L2 performance in the more complex task (+IR), the regression analyses (see Table 4) showed that the model for *syntactic complexity* (F (3, 44) = .856, p = .471) failed to reach a statistically significant level, implying that LP and WM could not predict syntactic complexity of the learners' performance in the more complex task. The regression model for *lexical complexity* reached a statistically significant level (F (3, 44) = 3.70, p = .019), explaining 20%

of the variance with explicit knowledge (p = .026) and WM (p = .011) making significant contributions to the model. Implicit knowledge, however, did not make a contribution (p = .325) to this model.

Table 4. Multiple regressions for LP and WM as predictors of performance in **+IR** task

Outcomes	**Predictors**	**Correlations**		**Regression models**			**Coefficients**				
DVs	**IVs**	***r***	***p***	***F***	***p***	***R²***	***B***	***SE***	***β***	***t***	***p***
Syntactic complexity	EXPLP	-.101	.246	.856	.471	.055	.12	.49	.48	.251	.803
	IMPLP	-.205	.081				-.86	.070	-.235	-1.21	.231
	WM	-.112	.224				-.154	.212	-.107	-.72	.470
Lexical complexity	EXPLP	**-.282**	**.026***	**3.70**	**.019***	**.201**	-.575	.260	-.393	-2.21	.032
	IMPLP	-.67	.325				.388	.375	.184	1.03	.306
	WM	**.328**	**.011***				2.62	1.13	.314	2.32	.025
Accuracy	EXPLP	**-.235**	**.050***	**2.97**	**.042***	**.168**	-.290	.443	-.119	-.655	.516
	IMPLP	**-.238**	**.050***				-.584	.639	-.166	-.914	.366
	WM	**.318**	**.014***				4.43	1.92	.317	2.30	.026
Speed fluency	EXPLP	-.216	.070	1.17	.331	.074	-1.78	1.33	-.256	-1.13	.188
	IMPLP	-.096	.258				.687	1.92	.068	.356	.723
	WM	.163	.134				6.18	5.80	.155	1.06	.293
Breakdown fluency	EXPLP	.128	.192	.511	.677	.034	.015	.183	.016	.080	.937
	IMPLP	.182	.107				.234	.265	.172	.882	.382
	WM	-.014	.321				-.089	.797	-.017	-.112	.921
Repair fluency	EXPLP	.135	.180	.373	.773	.025	.112	.112	.195	.995	.325
	IMPLP	.033	.413				-.078	.162	-.094	-.479	.634
	WM	-.042	.388				-.116	.488	-.035	-.237	.814

*$p < 0.05$, df (3, 44)

As for *accuracy*, the regression model was statistically significant (F (3, 44) = 2.97, p = .042), explaining 17% of the variance. WM (p = .014), explicit knowledge (p = .05), and implicit knowledge (p = .05) made significant contributions to the accuracy model. Regarding *speed fluency*, the regression model was not statistically significant (F (3, 44) = 1.17, p = .331),

revealing that speed fluency of the learners' performance could not be predicted by variations in their explicit or implicit LP or WM in this task.

To sum up, the results of the multiple regression analyses regarding +IR performance showed that the models for lexical complexity and accuracy were statistically significant, whereas the models for syntactic complexity and speed fluency failed to reach a statistically significant level. As for the individual contributions, WM and explicit L2 knowledge had significant contributions to both models of lexical complexity and accuracy. For lexical complexity, however, implicit L2 knowledge did not appear to make a significant contribution.

We conducted a post-hoc power analysis using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009) to examine the power of our t-tests. Although the significant results and effect sizes achieved in the study underline the effects of IR on task performance, the results of power analysis can reassure us about the strength of the findings. The power to detect a medium effect size of .5 (Plonsky & Oswald, 2014) was determined to be 0.96, and critical t (47) = 1.68. Running the analysis for a linear regression fixed model, we calculated the power of each individually significant regression model with an alpha level of 0.05 and a sample size of 48. The results showed a power of .78 for *accuracy*, and .85 for *lexical complexity.* The results of the power analysis suggest that although a reliable level of confidence could be maintained in the findings, the results should be interpreted with care.

Discussion

In this section, we first summarize the findings of the study and will then discuss the relationship between TC and task performance. We will also highlight the potential contribution of the findings to a more in-depth understanding of the relationship between TC and the two learner-internal factors of LP and WM. The main aims of this study were a) to investigate the effects of TC, operationalized in terms of the degree of IR required to complete the task, on

different aspects of performance, and b) to explore whether LP and WM mediated such effects in performance in these tasks. The results concerning the effects of TC replicated our previous findings (Awwad et al. 2017) in that performance in the +IR condition was associated with more accuracy, higher syntactic complexity in terms of subordination and length of AS unit, and less lexical complexity. Fluency in the +IR task was higher for speech rate, but not for other measures. Although the –IR task elicited longer clauses and longer pauses, these differences failed to reach a statistically meaningful level. These findings support the assumptions of Cognition Hypothesis only partially as an increase in accuracy of learners' performance is associated with an increase in syntactic complexity but not with higher lexical complexity. The results also partially support the predictions of Skehan's Limited Attentional Capacity as some aspects of fluency compete with accuracy and complexity, while other aspects increase with accuracy and complexity.

Research Questions 2 asked whether LP and WM mediated the effects of TC on performance. The results of the regression analyses suggested that variations in LP (both explicit and implicit knowledge) predicted up to 20% of the variance in accuracy of performance in both task conditions. Explicit L2 knowledge also predicted lexical complexity in the +IR task, suggesting that in a complex task, learners with a higher level of explicit L2 knowledge (and a stronger WM) produced more complex lexical items. For the model predicting speed fluency, while implicit and explicit L2 knowledge made a noticeable contribution to predicting speed fluency in the –IR task, the model did not reach a statistically significant level. The findings for models of fluency and lexical complexity imply that task design combined with learner-internal factors can explain variations in lexical complexity and speed fluency. Recent research in L2 fluency (Tavakoli, et al., 2017) has shown that while speed fluency is directly linked with LP, breakdown and repair fluency are to some extent linked to other personal non-proficiency related factors such as personal style.

The results also suggested that the learner-internal factors did not predict syntactic complexity of performance. While we had expected to see LP effects for all measures of performance, the non-significant results for syntactic complexity were rather surprising. In line with Gilabert and Munoz (2010), one conclusion we arrive at is that TC is potentially more crucial than learner-internal factors in determining which syntactic structures should be used during task performance. It is also possible to argue that the effects of TC override the advantages of a higher LP level.

As for the effects of WM, significant results were found only for accuracy across both task performances. This suggests that learners with a higher score in WM are likely to produce more accurate structures. We also found significant contribution of WM to lexical complexity in the +IR task performance. This result is in line with Gilabert and Munoz's (2010) finding in which WM effects on lexical complexity were reported. More central to the focus of the study was an examination of the possible interaction between TC, WM in task performance. The non-interaction effect between WM and TC for measures of syntactic complexity and fluency suggest that WM did not predict task performance. This finding is in line with Cho (2018), who observed no interactions between TC and WM. These results are, however, surprising when compared with previous research in SLA that considers WM as an important factor in L2 performance and acquisition (Ahmadian, 2015; Kormos & Trebits, 2011; Wen, 2015). While the findings of the current study imply that WM might play a different role in performance on tasks of varying TC, the results clearly show that more research is needed to examine the interaction between WM and TC.

While the internal-learner factors did not predict speed fluency in the +IR task, explicit and implicit knowledge made significant contributions to the models predicting speed fluency in the –IR task performance ($p < .01$ and $p < .04$ respectively). This finding implies that while fluency research (Tavakoli, et al., 2017) has shown a linear relationship between speed fluency

and LP, in a complex task speed fluency is influenced by factors other than proficiency, in this case TC.

The broader perspective to measuring LP in this study was expected to reveal differential contributions of implicit versus explicit knowledge to the models predicting performance. The results, however, indicated that when LP mediated the effects of TC, it usually involved a contribution from both types of knowledge. The only exception to this was the model for lexical complexity in +IR task in which only explicit LP helped predict task performance.

An important finding of the study is that TC demands and the language used to express such demands are inevitably and intricately linked. For example, the +IR task asked the participants to discuss the characters' intentions, to justify their actions and to predict the consequences of those actions. To address these requirements, most participants used hypothetical language of a formulaic nature (*I think*, *I suppose*), language of justification (*they want to xxx to; they are doing xxx because*) and linguistic units of prediction (*they are going to*). Such needs are likely to encourage the use of subordination and complex structures. On the other hand, some of these structures repeatedly used by the learners were of a formulaic nature, which enhances accuracy and fluency measured in a CALF framework (Boers et al., 2006). Therefore, although this task may add to the learners' cognitive load and perceptions of task difficulty, it has inevitably invited learners to use language that transmits intentions, predictions and justifications. The language requirements in a different cognitively demanding task might well encourage very different structures with rather different effects on measures of syntactic and lexical complexity.

As indicated in the results, +IR performances were associated with a higher ratio of subordination and longer AS units, but not with longer clauses. Several researchers (e.g. Awwad et al. 2017; Inoue, 2016; Skehan, 2014) have argued that mean length of clause seems to tap into a different aspect of syntactic complexity construct, and as such it is a useful measure to include in the analysis of speech performance. There is some emerging evidence in SLA

research (Abrams, 2019; Pallotti, 2009) to suggest that syntactic complexity is at least to some extent a function of individual and stylistic preference, and as such variance is inevitably anticipated among the speakers. The results reported in this study for syntactic complexity may potentially highlight presence of such variance among the speakers.

Interestingly, the performance in the +IR task was also associated with higher accuracy. That both measures of accuracy reached significant levels suggested that the language used under the +IR condition was more accurate. Once again, we interpret this result in relation to the use of formulaic language to express intentionality in the +IR task (e.g., *I think, I suppose, they want to*, etc.). As for lexical complexity, the lower lexical indices of performances in the +IR task suggest that narrating the +IR stories does demand the use of more varied and more sophisticated lexical items. This might also have been linked to the need to repeat the language of intentionality and prediction. It is possible to argue that while under –IR conditions, the learners were free to use various and perhaps more sophisticated words, while +IR tasks encouraged the learners to repeat certain words that could help them accomplish the task, i.e., explain the characters' thoughts, justify their actions and predict the consequences. As discussed earlier, our main aim in this study was not to map these findings with the CH or LAC models, but to move beyond these models and draw researchers' attention to the fact that our conceptualization and operationalization of TC is closely linked with the language needed to express TC demands. As for choice of analytic measures and their impact on the findings, we support Inoue's (2016) call for a more careful choice of analytic measures that are relevant to and useful for measuring task performance. From a methodological perspective, although recent research (Foster & Wigglesworth, 2016) has suggested that weighted clause ratio is a more sensitive measure of accuracy, we can see that the two global measures of accuracy, i.e., EFC and WCR, not only show very similar results, but also positively and strongly correlate (r = .86,

p = .000). For future research this implies that the use of one could to a large extent represent the other.

The results of the current study have significant implications for second language pedagogy. The results highlight the role of task design as a valuable pedagogic tool that can help promote opportunities for encouraging more accurate, complex and fluent language use. The findings, for example, imply that the +IR task provided a rich opportunity for use of hypothetical language, and language of justification and prediction. However, requirement of IR in a task encourages language of less lexical diversity and sophistication. From a pedagogic perspective, these variations can be effectively utilized in classroom to promote teaching and learning objectives. Another important pedagogic implication of these findings worth considering is the impact of TC on syntactic complexity. The fact that TC may override the power of LP in producing syntactically complex structures is an important finding to be taken into consideration in materials development and syllabus design. These results should encourage teachers to provide lower proficiency learners with cognitively demanding tasks that invite use of syntactically complex structures. Similarly, by a careful selection of task content, e.g., by choosing content that requires explaining different actions and/or justifying them, teachers can inspire leaners to aim for more complex language both syntactically and lexically.

Conclusions

The significance of the findings of the current study lies in its contribution to issues related to understanding the role of LP and WM in L2 performance when TC is manipulated along the IR continuum. The present study engaged with disciplinary debates about the impact of TC on second language performance and whether individual differences of WM and LP mediate such effects. Building upon previous research, the findings have extended our understanding of a less-researched aspect of TC, i.e., varying levels of IR demands, by providing a detailed definition and a careful operationalization of the IR construct. Our study provides a unique

contribution to the literature by exploring task performance and TC in relation to different levels of WM and LP. In line with previous research (e.g., Baralt, 2015; Kormos & Trebits, 2011; Malicka & Levkina, 2012), our results imply that TC has a substantial influence on task performance. Using regression analysis, the results showed that LP and WM predicted accuracy in both task types and lexical complexity in the +IR task. The results also suggested that LP predicted speed fluency, but only in the –IR task, implying that the speed of performance can be predicted, at least to some extent, by the learners' LP if a task is not cognitively demanding. However, syntactic complexity of learner performance cannot be explained by levels LP or WM. The findings also suggest that TC, at least the way it is operationalized in the study, determines the linguistic units that will emerge in task performance. Therefore, TC and the language that communicates the cognitive demands of a task inherently interact with one another. These results also imply that examining TC in isolation may provide a too simplistic picture of the processes involved in L2 production and acquisition.

References

Abrams, Z. I. (2019). The effects of integrated writing on linguistic complexity in L2 writing and task-complexity. *System*, 81, 110-121.

Ahmadian, M. (2015). Working memory, online planning and self-repaire bahaviour. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (Vol. 87, pp. 160-174). Bristol: Multilingual Matters.

Allan, D. (2004). *Oxford Placement Test. University of Cambridge Local Examination Syndicate.* Oxford: Oxford University Press.

Awwad, A., Tavakoli, P., & Wright, C. (2017). "I think that's what he's doing": Effects of intentional reasoning on second language (L2) speech performance. *System*, 67, 158-169.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford university press.

Baralt, M. (2015). Working memory capacity, cognitive complexity and L2 recasts in online language teaching. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 248-269). Bristol: Multilingual Matters.

Beneš, L., & Jiránek, V. (1976). Pat & Mat, Czech stop-motion animated series. Retrieved from http://en.patmat.cz/home-pat-and-mat/

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language teaching research*, 10(3), 245-261.

Boersma, P., & Weenink, D. (2008). Doing phonetics by computer: Praat: ver 4.5.01 [Computer program]. *Computer software*, http://www.fon.hum.uva.nl/praat/.

Breck, E. (1998). SoundScriber. Michigan: University of Michigan. Retrieved from http://www-personal.umich.edu/~ebreck/code/sscriber/

CEFR. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment. Council of Europe.* Cambridge, UK: Cambridge University Press.

Cho, M. (2018). Task complexity, modality, and working memory in L2 task performance. *System*, 72, 85-98.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review, 12*(5), 769-786.

de Jong, N., & Vercellotti, M. (2015). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research, 20*(3), 387-404.

Declerck, M., & Kormos, J. (2012). The effect of dual task demands and proficiency on second language speech production. *Bilingualism: Language and cognition, 15*(04), 782-796.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language learning* (pp. 313-348). Oxford: Blackwell.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 119-138). Oxford: Blackwell

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42, pp. 3-25). Bristol: Multilingual Matters.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464-491.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4), 1149-1160.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Foster, P. & Tavakoli, P (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning*.59(4): 866-896.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics, 36*, 98-116.

Fulcher, G. (2014). *Testing second language speaking*. London: Routledge.

Gaillard, S., & Tremblay, A. (2016). Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. *Language Learning, 66*(2), 419-447.

Gilabert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: The role of working memory capacity. *International Journal of English Studies, 10*(1), 19-42.

Gilhooly, K. (2004). Working memory and reasoning. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 49-77). Cambridge: Cambridge University Press.

Graesser, A., McNamara, D., & Louwerse, M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. Sweet & C. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford Publications.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. *Studies in second language acquisition*, 14(1), 25-38.

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning. *Studies in Second Language Acquisition*, 27, 129–140.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal, 44*(4), 487-505.

Ishikawa, T. (2006). The effect of task complexity and language proficiency on task-based language performance. *The Journal of AsiaTEFL, 3*(4), 193-225.

Ishikawa, T. (2008). The effect of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL, 5*(1), 29-63.

Jackson, D., & Suethanapornkul, S. (2013). The Cognition Hypothesis: A Synthesis and Meta-Analysis of Research on Second Language Task Complexity. *Language Learning, 63*(2), 330-367.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809-854.

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 39-60). Amsterdam: John Benjamins.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 267-285). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*(1), 48-60.

Kuiken, F., & Vedder, I. (2011). Task complexity and linguistic performance in L2 writing and speaking. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 91-104). Amsterdam: John Benjamins.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14(4), 389-433.

Leal, T. (2018). Data analysis and sampling. In A. Gudmestad & A. Edmonds (Eds.), *Critical Reflections on Data in Second Language Acquisition* (pp. 51-63). Amsterdam: John Benjamins.

Leighton, J. (2004). Defining and describing reason. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 3-11). Cambridge: Cambridge University Press.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*(3), 387-417.

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge MA: MIT Press.

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553-568.

Malicka, A. (2014). The role of task sequencing in monologic oral production. In P. Robinson (Ed.), *Task sequencing and instructed second language learning* (pp. 71-93). London: Bloomsbury.

Malicka, A., & Levkina, M. (2012). Measuring task complexity: does L2 proficiency matter. In A. Shehadeh & C. Coombe (Eds.), *Task-based Language Teaching in Foreign Language Contexts: Research and Implementation* (pp. 43-66). Amsterdam: John Benjamins.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language testing*, *19*(1), 85-104.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 5-19.

Meara, P., & Miralpeix, I. (2016). *Tools for Researching Vocabulary*. Bristol, UK: Multilingual Matters.#

Mitchell, A. E., Jarvis, S., O'Malley, M., & Konstantinova, I. (2015). Working memory measures and L2 proficiency. In Z. Wen, M. Borges & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (Vol. 87, pp. 270-283). Bristol: Multilingual Matters.

Mota, M. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos*, *24*, 69-104.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578.

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, 19(4), 218-233.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590-601.

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom- based study. *The Modern Language Journal*, *95*(1), 162-181.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27-57.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching*, *45*(3), 193-213.

Robinson, P. (2011). *Second language task complexity: researching the cognition hypothesis of language learning and performance* (Vol. 2). Amsterdam: John Benjamins.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 87-121). Amsterdam: John Benjamins.

Sasayama, S. (2016). Is a 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, 100(1), 231-254.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language teaching*, *36*(01), 1-14.

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 1-26). Amsterdam: John Benjamins.

Skehan, P. (2015). Limited Attention Capacity and Cognition. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 123-155). Amsterdam: John Benjamins Publishing.

Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 199-220). Amsterdam: John Benjamins.

Tavakoli, P. & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*. 58(2): 439-473.

Tavakoli, P. and Skehan, (2005). P. Strategic planning, task structure and performance testing. In R. Ellis (Ed.). *Planning and task performance in a second language* (pp. 239-277). Amsterdam: John Benjamins.

Tavakoli, P. Campbell, C. & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly*. 50(2): 447-471.

Tavakoli, P. Nakatsuhara, F. & Hunter, A-M. (2017). Scoring validity of the Aptis Speaking test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online*. ISSN 2057-5203 London: British Council.

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The Effects of Mode and Task Complexity on Second Language Production. *Language Learning*, 67(2), 394-430.

Wang, Z., & Skehan, P. (2014). Structure, lexis, and time perspective. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 155-185). Amsterdam: John Benjamins.

Wen, Z. (2015). Working memory in second language acquisition and processing: The phonological/executive model. In Z. Wen, M. Mota & A. McNeill (Eds.), *Working Memory in Second Language Acquisition and Processing* (pp. 41-62). Bristol: Multilingual Matters.

Wen, Z., Mota, M., & McNeill, A. (2015). *Working memory in second language acquisition and processing* (Vol. 87). Bristol: Multilingual Matters.

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, *46*(4), 680-704.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2015). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, 33(4), 497-528.